DOCUMENT RESUME

ED 097 376 TM 004 018

AUTHOR	Jaeger, Richard M.
TITLE	A Primer on Sampling for Statewide Assessment.
INSTITUTION	Educational Testing Service, Princeton, N.J. Center for Statewide Educational Assessment.
PUB DATE	73
NOTE	60p.
AVAILABLE FROM	Center for Educational Assessment, Educational Testing Service, Princeton, N.J. 08540 (Free)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Definitions; *Educational Assessment; *Guides; Objectives; *Sampling; *State Surveys

ABSTRACT

This paper is a primer on sampling procedures for statewide assessment. The careful reader should gain substantial knowledge about the promises and pitfalls of sampling for assessment. The primer has three basic objectives: (1) to define terms and concepts basic to sampling theory and its application, including population, sampling unit, sampling frame, probability sampling procedures, estimate, population parameter and estimator, estimator bias, variance, mean square error and efficiency, and consistency; (2) to illustrate some of the ways sampling procedures can be used to achieve realistic assessment objectives; and, (3) to describe issues that arise when sampling procedures are used, and the factors that contribute to their resolution. Objectives two and three include discussions of simple random sampling, stratified random sampling, systematic sampling, cluster sampling, and matrix sampling. The appendix gives an example of an evaluation of alternative cluster sampling procedures. (SE)

CSEA

ED 097376

BEST COPY AVAILABLE

A Primer on Sampling for Statewide Assessment

Richard M. Jaeger

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
DOROTHY URBAN
EDUCATIONAL TESTING SERVICE
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

TM 004 018

CENTER FOR STATEWIDE EDUCATIONAL ASSESSMENT
EDUCATIONAL TESTING SERVICE • PRINCETON, NEW JERSEY

BEST COPY AVAILABLE

Copyright © 1973 by Educational Testing Service

Educational Testing Service is an Equal Opportunity Employer

ED 097376

A PRIMER ON SAMPLING FOR STATEWIDE ASSESSMENT

Published by the Center for Statewide Educational Assessment which is supported by funds from the Ford Foundation.

Richard M. Jaeger
Visiting Research Psychologist
Educational Testing Service

TM 004 018

PREFACE

This paper is a brief introduction to finite population sampling methods, specially prepared for those concerned with statewide assessment programs. The sampling procedures described in the paper are those most likely to be useful in achieving the objectives of statewide assessment.

The paper is intentionally non-mathematical. While it presumes knowledge of the fundamental concepts of statistical inference, it does not require any prior exposure to the formalities of sampling. All sampling terms used in the paper are carefully defined. Descriptions of sampling procedures make use of these definitions, and avoid unnecessary technicalities. The paper is intended to be a resource for those engaged in the practice of statewide assessment, and makes no claim to comprehensiveness as a theoretical treatise.

Helpful suggestions and clarifications of some otherwise opaque issues were provided by Nancy Bruno, Paul Campbell, Henry Dyer and Robert Linn. I want to express my appreciation for their careful reviews of early drafts. I am solely responsible for any remaining inaccuracies.

Princeton, New Jersey

Richard M. Jaeger

TABLE OF CONTENTS

About this Paper.....1

Some Terms and Concepts.....2

Population.....2

Sampling Unit.....4

Sampling Frame.....5

Probability Sampling Procedures.....6

Estimate, Population Parameter and Estimator.....8

Estimator Bias.....9

Variance, Mean Square Error and Efficiency.....12

Consistency.....17

Using Sampling in Statewide Assessment.....20

Simple Random Sampling.....22

Stratified Random Sampling.....25

Systematic Sampling.....30

Cluster Sampling.....34

Matrix Sampling.....44

Summary.....48

References.....50

Appendix A: Evaluation of Alternative Cluster Sampling Procedures--An example.....51

A Primer on Sampling for Statewide Assessment

About this Paper

When a statewide assessment is planned, one of the first issues that arises is who should be tested? Even after a state has decided to test students in certain grades or at certain age-levels, the question, who should be tested?, remains. Should all fourth-graders be tested, or should some be selected for testing?

In some states, the objectives and purposes that give rise to assessment include a desire to secure test results for each student in a grade; the assessment goals include individual assessment as well as institutional assessment. When individual assessment is desired, the "who to test" question is answered by the selection of a grade or age-level for assessment. When individual measurement is not a goal of statewide assessment, it is usually economical and administratively desirable to select a sample of students for testing, rather than testing all students.

This paper is intended to be a primer on sampling for statewide assessment. If its purpose is achieved, the careful reader will gain substantial knowledge about the promises and pitfalls of sampling for assessment. The reader will not become an instant sampling expert; no short paper can accomplish that goal. Instead, the dedicated reader will become a "sampling conversationalist", able to meet a sampling expert at least half way, and able to knowledgeably discuss sampling issues important to his state's assessment. Further, he will be able to converse in the language of the expert.

The goal of creating "sampling conversationalists" will be pursued in three ways:

1) By defining terms and concepts basic to sampling theory and its applications;
2) by illustrating some of the ways sampling procedures can be used to achieve realistic assessment objectives; and
3) by describing issues that arise when sampling procedures are used, and the factors that contribute to their resolution.

The balance of this paper is in two parts. The first part provides definitions of some of the most important terms and concepts fundamental to the language of sampling. In the second, consideration is given to two potential objectives of a statewide assessment, and the ways various sampling procedures can contribute to their achievement. In part two, the reader is faced with alternatives and choices, and then presented with facts to help him make decisions.

Some Terms and Concepts

Population

In any sampling study, there is a definable group or aggregation of elements from which samples are selected. This aggregation of elements is called the population of the study. Technically, any aggregation of elements that have at least one attribute in common can form a population. In a statewide assessment, some examples of populations that might be of interest are all public schools in the state that enroll sixth-graders, all sixth-graders enrolled in public schools in the state and all public-school sixth-graders in the state who are children of migrant agricultural workers. From these examples, it is clear that populations can be composed

of individuals or institutions. Similarly, populations can be composed of people or things. The first population, all public schools in the state that enroll sixth-graders, is defined by two attributes: control of school (public) and grade-level offerings (sixth grade); the second population is also defined by two attributes: grade-level and public-school enrollment; the third population has three defining attributes: grade-level, public-school enrollment, and parental occupation.

These examples of populations have some important characteristics in common. Each is composed of a finite number of elements (sixth-graders in the state, schools with sixth-graders in the state, etc.), and each is defined by attributes that are easily recognized. That is, one can easily decide whether an element is or is not a member of the population.

Some populations that are infinite in size may be encountered in a statewide assessment. An example of an infinite population is "all multiple-choice test items that could ever be written, that purport to measure reading comprehension". In contrast to the first examples, this population is not defined by attributes that are easily recognized. If faced with a test item that contained a paragraph of prose followed by four questions on the main theme of the paragraph, most of us would say that the item was a "reading comprehension" item, and therefore a member of the population. But what about an arithmetic word problem..."If it took six men five days to dig a ditch...?". Clearly, reading comprehension is a skill required to answer the item correctly. Yet it requires more than reading comprehension to compute a correct solution. Is the item a member of the population? The answer is debatable.

All of the sampling procedures discussed in this paper assume that the populations to be sampled are finite. This is a realistic assumption whenever students, classes, schools or school districts are sampled. Unlike finite populations, infinite populations are somewhat intangible and exist only in the mind of the beholder. However, there is a well-developed theory of sampling from infinite populations, so they present no insurmountable statistical problems.

Another way of defining a population is "the aggregation of elements that is of central interest in a study". This is an admittedly loose definition that might upset some statistical purists, but it helps to point out the practical significance of populations. In a real-world study such as a statewide assessment, populations are not theoretically-defined entities that exist for the fascination of statisticians; they are the central focus of the study. For example, in your statewide assessment you may want to know the proportion of public-school fourth-graders whose reading comprehension score is below the 25th percentile on a national norm distribution. Here, the population of interest is all fourth-graders enrolled in the public schools of your state. The population is real, and of practical interest. If you test every public-school fourth-grader in the state, you can determine the proportion exactly (provided there are no missing data, all absentees are tested at a later date, etc.).

Sampling Unit

Populations are made up of elements termed sampling units. The sampling units into which the population is divided must be unique, in the sense that they do not overlap, and must, when aggregated, define the whole of the pop-

ulation of interest. Sampling units that might be used in statewide assessments include students, class-sections, homerooms, teachers, schools, and school districts. These examples of sampling units clearly define unique elements (one student is different from another; schools that have the same grade-levels are generally unique units) that can be readily counted and aggregated.

The definitions given for "population" and "sampling unit" may appear to be circular. But perhaps that's as it should be, since sampling units, when aggregated, make up a population, and a population is an aggregation of sampling units.

Sampling Frame

When "selecting a sample", one is in fact selecting sampling units from the aggregation that composes the population. For a unit to be selected, it must be identifiable. A list that uniquely identifies all of the units in a finite population is termed a sampling frame. A sampling frame for statewide assessment might consist of a list of all schools in the state that enroll pupils in grades one through six, or a list of all secondary students enrolled full time in vocational education programs.

When assembling a sampling frame, care must be taken to ensure that it corresponds precisely to the population of interest. In the first example above, a sampling frame that consists of all schools in the state that enroll pupils in grades one through six would be composed of non-public schools as well as public schools. If the population of interest consisted only of public elementary schools, this sampling frame would be inappropriate. First, non-public schools would be listed in the frame although

they are not elements of the population of interest. The erroneous listing of elements outside the population of interest is known as "overregistration". Second, the definition of an "elementary school" differs from state to state. In some states, a school is classified as an elementary school if it enrolls pupils in any grade between kindergarten and grade six. In other states, an elementary school is defined as a school that enrolls pupils in any grade between kindergarten and grade eight. In states with the latter definition, there may be schools that enroll only seventh and eighth-graders, that would be elements of a population of elementary schools. Yet these schools would be excluded from a sampling frame that listed schools with pupils in grades one through six. In this case, elements of the population of interest (all public elementary schools) would be excluded from the sampling frame (all schools that enroll pupils in grades one through six). This type of error in constructing a sampling frame is known as "underregistration".

The point to be made is that populations of interest in statewide assessment should be clearly and precisely defined. Then sampling frames that include only elements in the populations of interest, and all elements in the populations of interest, should be carefully constructed.

Probability Sampling Procedures

When sampling is used in statewide assessments, the financial objectives are clear. The desire is to save money and time by measuring or testing only a sample of students, yet be able to make accurate statements about a population of students. Probability sampling procedures often allow these objectives to be achieved, and in addition, allow one to determine the likelihood of making inaccurate statements about a population.

Probability sampling procedures have three characteristics in common. First, the procedures are applied to populations where the units which compose the population and the units which are excluded from the population are explicitly defined. That is, given a potential sampling unit, one can say unequivocally whether it is in the population or not. Second, the chances (or probability) of selecting any potential sample can be specified. Third, every sampling unit in the population has a positive chance of being selected. It isn't necessary that every potential sample have an equal chance of being selected, just that the chance of selecting any potential sample can be specified.

The formal definition of a probability sampling procedure might appear somewhat formidable, and perhaps unenlightening as well. Sometimes even simple things are obscured by formality (a square is a right parallelopiped composed of four pairwise orthogonal line segments...). Instead of pursuing the definition further, consider some sampling methods that are <u>not</u> probability sampling procedures. Assume that an assessment objective is to determine the average social studies achievement of eighth-graders in each school district in the state. Suppose that a particularly large school district decides to test eighth-graders in half its schools and use their average achievement as an estimate of the average for all eighth-graders. Suppose they decide to select for testing, those schools that are closest to the district research office. With this plan, they'll select the school closest to the research office first, the second closest school second, and so on, until half the schools in the district have been "sampled". This isn't

a probability sampling procedure, because it violates the third characteristic of such procedures. All the schools with eighth-graders that are farthest from the district research office are contained in the sampling frame, but they don't have any chance (zero probability) of being selected. This same violation would occur with any sampling procedure that selects schools only from a prescribed section of the district.

These sampling procedures cause problems not because they violate an arbitrary rule, but because they are likely to produce samples that don't represent the population. The district research office is probably in the older or downtown area of the system. Schools near it are more likely to enroll students from lower socio-economic status families than in the district as a whole, and the achievement of these students is therefore likely to be lower than in the district as a whole. So again, the rules are not just statistical artifacts. They help to prevent trouble in the practical world of assessment.

Estimate, Population Parameter, and Estimator

In addition to providing procedures for collecting data, sampling theory provides formulas for estimating characteristics of populations, such as averages, proportions, and totals. When a sample is drawn from a population, and a statistic (such as an average) is computed from data on the units sampled, the number that results is called an estimate. For example, if it is found that a sample of ten students selected from a population of 200 has an average arithmetic score of 42, the number 42 is an estimate of of the average for the entire population of 200. The average for the entire

population would be an example of a population parameter. In general, population parameters are unknown characteristics of populations that survey researchers would like to know. If every element in a population is measured, the value of the population parameter can be determined. Instead of measuring every population element, a survey researcher will measure only elements in a sample and, from these data, compute an estimate of the population parameter. Formulas that are used to compute estimates from sample data are termed estimators.

In a statewide assessment, the average educational level of teachers in the state might be estimated by sending a questionnaire to a sample of teachers, and computing an average for the sampled teachers. An average computed from the questionnaire responses of the sample is an estimate, and a formula used to compute the average for the sample of teachers is an estimator.

Estimator Bias

When a population is finite, the number of different samples that can be drawn from it is also finite. A list can be made for any finite population, containing all of the samples of a given size that could possibly be drawn from it. For example, suppose that a school district has four high schools and an assessment director wants to sample two of the four. If the schools are numbered from one to four, the six different samples of two schools that could be drawn are as follows:

Sample	Schools in Sample
A	1, 2
B	1, 3
C	1, 4
D	2, 3
E	2, 4
F	3, 4

Suppose the assessment director wants to know the average number of certified science teachers per high school in the district, and decides to estimate the average by collecting data in two of the four schools. In this example, the population parameter is the actual average per school for the four schools in the district. Data from each sample would provide an estimate of this population parameter, and since six different samples could be selected, six different estimates are possible.

Continuing the example, suppose that an estimate of the population average per school was actually calculated using data from each sample, and the six estimates were then tabulated. It would then be possible to calculate the average of these six estimates. If the average value of the estimates was equal to the population average, the estimator (formula used to calculate each estimate) would be termed an unbiased estimator. If, on the other hand, the average of the sample estimates was either larger or smaller than the population average, the estimator would be biased.

In general, an estimator is said to be biased if the average of the estimates it would produce (if the average were to be taken over all possible samples of a given size) were either larger or smaller than the population parameter. If the average of all estimates were to equal the population parameter, the estimator would be termed unbiased.

It should be intuitively clear that unbiased estimators are desirable. An assessment director would be happiest if every estimate computed from a sample was equal to the population parameter of interest. Since this utopian condition will hardly ever be true, it is at least nice to have

the average of the estimates equal the population parameter.

Although unbiased estimators are desirable, a biased estimator can sometimes be useful if the magnitude of the bias (the difference between the average estimate and the population parameter) is small. Under some conditions likely to be encountered in a statewide assessment an unbiased estimator may actually be rejected in favor of a biased one.

At this point, the reader may wonder how estimator bias can be computed using data from a single sample. The answer is, that it can't be computed from sample data. To compute bias, one would have to know the value of the population parameter. If the population parameter were known, there would be no reason to sample.

The bias (or lack of bias) of a sampling and estimation procedure is actually determined from the estimator used (a mathematical formula), and the mathematical assumptions that underlie the sampling procedure. Determination of bias is an algebraic procedure that doesn't depend upon data at all (Murthy, 1967; Cochran, 1963).

NUMERICAL EXAMPLE. Suppose that the average number of certified science teachers per school was known to be equal to 3.5 for the four schools in the district, and the estimates computed for the six possible samples were as follows:

Sample	*Schools in Sample*	*Estimate*
A	*1, 2*	*4.3*
B	*1, 3*	*3.2*
C	*1, 4*	*2.8*
D	*2, 3*	*3.7*
E	*2, 4*	*3.2*
F	*3, 4*	*3.9*
	Total	*21.1*

The average of the six estimates would equal

$$\frac{21.1}{6} = 3.52.$$

The estimator used would then be slightly biased, since the true value of the population parameter is 3.50, and the average of the estimates produced by all possible samples of size two is 3.52. The magnitude of the bias is equal to the difference between the population parameter value, and the average of the six estimates: 3.50-3.52 = -0.02.

**In this numerical example and in those that follow, hypothetical data are used. It is critically important to recognize that these examples have been constructed solely to illustrate the definitions of sampling concepts presented in the main body of the paper. Each example assumes a situation that is totally fictitious, and unlike the situations that will be encountered in practice. Namely, it is always assumed that the values of population parameters are known, and that estimates are available for all of the samples that could possibly be selected.*

In a practical sampling situation, population parameters will not be known. (If they were known, sampling would be unnecessary). Additionally, only one sample will be selected, and only one estimate of the population parameter will be computed. The variance of the sample estimate (see the following section of the text) will not be directly computable from the data provided by a single sample. However, the variance of the sample estimate can almost always be estimated from the data provided by a single sample, and this estimate will almost always be computed in practice.

Variance, Mean Square Error and Efficiency

When an estimate of a population parameter is computed, it will rarely be equal to the population parameter. The difference between the estimate and the population parameter is known as an error of estimation. In the

numerical example of the last section, the average number of certified science teachers per school was assumed to be equal to 3.5 for the four schools in the district, and the estimate computed from Sample F was assumed to be 3.9. With these assumptions, the error of estimate would be (3.5) - (3.9) or -0.4.

If an estimator is unbiased, its variance is equal to the average of the squared errors of estimate, when the average is computed over all possible samples of a given size. Suppose that the estimator in the example of the last section had been unbiased. Then applying this formula for variance, the error of estimation would be computed for each of the six sample estimates, each of these would be squared, and the average of the six squared errors would equal the variance.

For a given sampling procedure and samples of a given size, the most desirable unbiased estimator is the one with the smallest variance. The smaller the variance of an unbiased estimator, the smaller the chance that a large estimation error can occur.

When an estimator is biased, its variance is also defined as the average of squares of differences. But instead of squaring the difference between each estimate and the population parameter, the variance of a biased estimator requires that the difference between each estimate and the average of all estimates be squared. The average of the squares of these differences is taken over all potential samples of a given size.

NUMERICAL EXAMPLE. Consider once again the hypothetical data presented in the last numerical example. In that example, the average number of certified science teachers per school was assumed to equal 3.5, in a school district

with four schools. All possible samples of two schools were identified, and estimates of the average number of certified science teachers per school were assumed to be as follows:

Sample	*Estimate*
A	*4.3*
B	*3.2*
C	*2.8*
D	*3.7*
E	*3.2*
F	*3.9*

The average of these estimates was found to equal 3.52. These data may now be used to compute the variance of the estimator:

Sample	*Estimate*	*Difference Between Estimate and Average*	*Square of Difference*
A	*4.3*	*4.3-3.52 = 0.78*	*0.6084*
B	*3.2*	*3.2-3.52 =-0.32*	*0.1024*
C	*2.8*	*2.8-3.52 =-0.72*	*0.5184*
D	*3.7*	*3.7-3.52 = 0.18*	*0.0324*
E	*3.2*	*3.2-3.52 =-0.32*	*0.1024*
F	*3.9*	*3.9-3.52 = 0.38*	*0.1444*
		Sum of Squares:	*1.5084*

Variance of Estimator = (1.5084)/(6) = 0.2514

The definitions of variance for biased estimators and unbiased estimators are illustrated by Figures 1A and 1B, below. Each figure shows a distribution of estimates across all potential samples from a population.

BEST COPY AVAILABLE

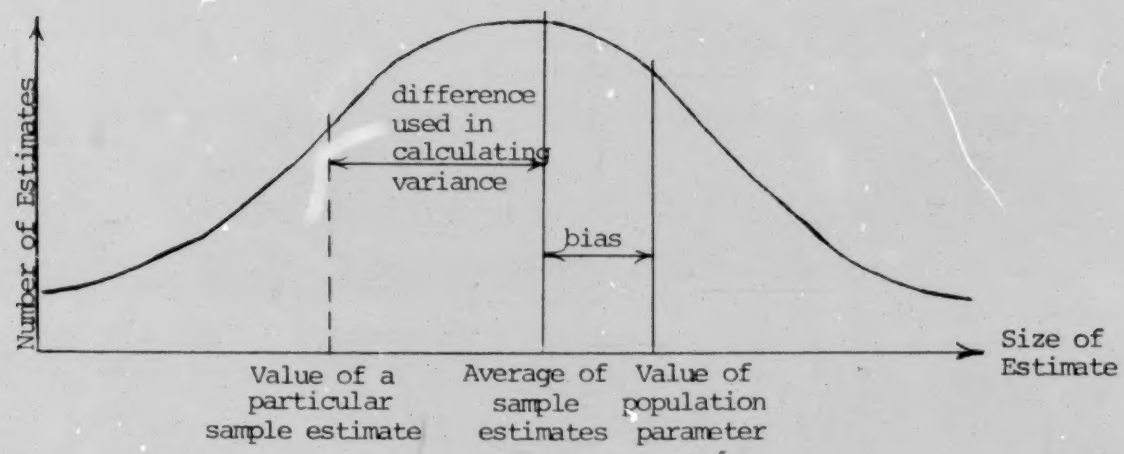

Figure 1A: Distribution of estimates for a biased estimator

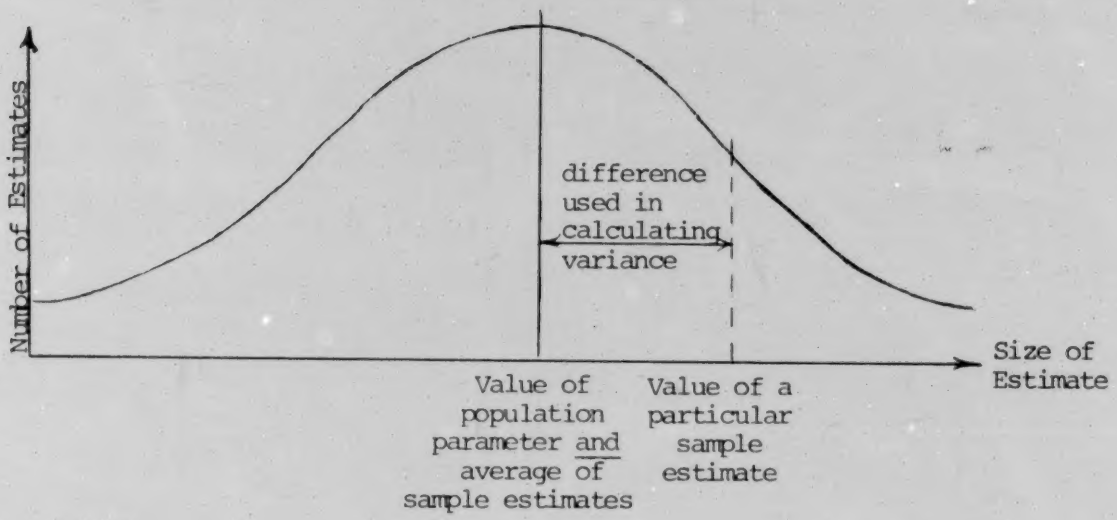
Number of Estimates
difference used in calculating variance
Size of Estimate
Value of population parameter and average of sample estimates
Value of a particular sample estimate

Figure 1B: Distribution of estimates for an unbiased estimator

In Figure 1A, the average of all estimates and the population parameter have different values, and the difference between them is equal to the bias of the estimator. In Figure 1B, the average of all estimates and the population parameter have the same value, since the estimator is unbiased.

If an assessment director has a choice of using two unbiased estimators, the one with the smallest variance should be selected. But what if the choice is between a biased estimator and an unbiased estimator? The biased estimator may have the smallest variance but its bias may be large, and the proper choice is unclear. The assessment director needs some way of comparing the magnitude of estimation errors of biased and unbiased estimators. A useful measure for this purpose is called the mean square error. Mean square error equals the sum of the estimator variance and the square of the estimator bias,

$$\text{Mean square error} = \text{Variance} + (\text{Bias})^2.$$

NUMERICAL EXAMPLE: Using the data of the previous numerical examples in the formula for the mean square error,

$$\text{Mean Square Error} = 0.2514 + (-0.02)^2$$
$$= 0.2514 + 0.0004$$
$$= 0.2518$$

In this numerical example, the mean square error of the estimator is clearly dominated by the variance. Although the estimator is biased, the magnitude of the bias is very small, and bias contributes an insignificant amount to the mean square error.

For an unbiased estimator, the mean square error and the variance are equal, since the bias is zero.

For a given sample size, an estimator that has a smaller mean square error than another is said to be more efficient. For a given sampling procedure, the most efficient estimator should always be used, since it will provide the smallest estimation errors, on the average. When different sampling procedures are used, a less efficient estimator may be preferred if its sampling procedure is less costly or more convenient. In the practical world of statewide assessment, it may be worthwhile to take a larger sample if the sampling procedure that can be used is more administratively convenient or less expensive to complete.

Consistency

Some amount of error in the estimation of population parameters from sample data is almost inevitable. However, the magnitude of errors likely to occur can often be controlled. With some sampling and estimation procedures, the mean square error value can be reduced by drawing larger and larger samples, and estimation error is reduced to zero when the sample size equals the population size. Such procedures are said to provide consistent estimation. A sampling and estimation procedure is said to be inconsistent if sampling errors can occur even when the sample size equals the population size.

When lack of consistency is encountered in practice, the sampling is usually being done "with replacement". In a "with replacement" procedure, an element of a population can enter the same sample more than once. Although lack of consistency can occur when elements are sampled without replacement (once an element is sampled it is removed from the population), it is not encountered in practical problems.

As an example of a "with replacement" sampling procedure, consider the case discussed in conjunction with estimator bias, above. In that example, two schools were sampled from a population of four schools. If sampling were to be done with replacement, ten different samples of two schools could be drawn. In addition to the six samples listed in the previous example, the following are possibilities:

Sample	Schools in Sample
G	1, 1
H	2, 2
I	3, 3
J	4, 4

More to the point, one could select many different samples of four schools, such as:

Sample	Schools in Sample
A	1, 2, 3, 3
B	1, 1, 2, 3
C	1, 1, 3, 4
D	1, 1, 1, 1
E	3, 4, 4, 4

Unless the number of certified science teachers was the same in all schools, each of these samples would provide a different estimate of the average number of science teachers per school. As a result, sampling errors could occur even though the sample size and the population size were the same.

Lack of consistency becomes a problem of real concern in two situations. First, when the mean square error of an estimator is not reduced in size in some orderly way, as the sample size is made larger and larger. Second, when the size of the sample necessary to achieve an acceptable mean square error is close to the size of the population. Several sampling and estimation procedures that are otherwise attractive for statewide assessment may produce these problems in some situations. These procedures, and the potentially problematic conditions, are described in the next part of this paper.

NUMERICAL EXAMPLE. Consider once again the hypothetical situation described in previous numerical examples, but suppose that a "with replacement" sampling procedure is used. Assume that all samples of size one, two, three, and four schools are selected, and the mean square error of the estimator is computed for each sample size. Suppose that the results are as follows:

Sample Size	*Mean Square Error*
1	*1.25*
2	*0.64*
3	*0.88*
4	*0.22*

This example illustrates two kinds of inconsistency. First, the mean square error does not become progressively smaller as the sample size is increased; the mean square error for samples of three schools is larger than the mean

square error for samples of two schools. Second, the mean square error is larger than zero for samples of four schools, even though there are only four schools in the population.

Clearly, the first kind of inconsistency is intolerable. A sampling researcher never knows how large the mean square error will be, although it can be estimated for many sampling procedures. Unless estimates are made for every possible sample size (which is sometimes impossible), the researcher can't determine an appropriate sample size with any degree of confidence; a large sample may be less efficient than a small sample.

Using Sampling in Statewide Assessment

Whether sampling is useful for statewide assessment depends primarily on the objectives of the assessment, and secondarily on the capabilities of those conducting the assessment. For some assessment purposes, usually when assessment results are desired for individual students, sampling will not be useful at all. For other purposes, as when assessment results are desired for individual classrooms, sampling may be feasible but impractical. But for many assessment purposes, sampling will not only be feasible, but a practical route to saving time, dollars and effort.

The capabilities of the agency conducting the assessment have been deemed secondary when considering the usefulness of sampling, since considerable help--through consultants or outside agencies--is likely to be

readily available. Further, the costs of such assistance are likely to be more than repaid through the savings afforded by sampling.

Some sampling procedures are both feasible and practical for some assessment purposes, but infeasible or impractical for others. For example, simple random sampling (which is discussed below) may be impractical for determining the average achievement of pupils in a particular grade throughout a state (the impracticality stems from the need for a single list of all pupils enrolled throughout the state), but practical and feasible for determining the average achievement of pupils in a particular grade in each school in the state. In the latter case, separate simple random samples might be selected from each school, using readily-available lists in each school district.

To this point, this paper has been concerned with the language of sampling--basic terms and concepts necessary to an understanding of sampling and samplers. We shall now change course by considering two practical assessment objectives gleaned from actual state assessment reports, and describing how sampling procedures could be used in achieving these objectives.

Objective 1: Determining the Average Reading Achievement of all Fifth-Graders in the State.

An obvious way of determining the average reading achievement of all fifth-grade pupils in a state is to test them all, record their scores, and compute the average. This procedure, known as taking a census of fifth-graders, was actually followed in the state that reported this objective.

For many objectives, and particularly when estimating statewide averages, taking a census is wasteful and unnecessary.

Simple Random Sampling

One procedure that could be used to achieve Objective 1 is called simple random sampling; a procedure in which every potential sample has an equal chance of being selected. Merely computing the arithmetic average of data from a simple random sample will provide an estimate of the population average. This sampling and estimation procedure is unbiased and consistent, and there are well-known formulas for estimating the mean square error of the sample average (Hansen, Hurwitz and Madow, 1953).

To estimate the average reading achievement of fifth-graders in a state through simple random sampling, the procedure would be as follows. First, a sampling frame would be constructed by listing each fifth-grader enrolled in the state, and assigning a unique number to each listed pupil. The sampling frame would include all enrolled fifth-graders or only fifth-graders enrolled in public schools, depending on the population of interest. Once the sampling frame was constructed, a table of random numbers would be used to select a sample of the desired size. A number would be drawn from the random number table, and the pupil with the corresponding number would be added to the sample. If a number drawn from the table either exceeded the largest number on the list of pupils, or repeated a number already drawn, it would be discarded. Selection of random numbers from

the table and corresponding pupils from the list would continue, until the desired sample size was reached.

A practical problem that we have skirted so far will arise time and time again in sampling. Just what is the "desired sample size" and how can it be determined? With simple random sampling, the desired sample size can be computed through straightforward application of a formula given by Hansen, Hurwitz and Madow (1953), Cochran (1963) or in many other books on sampling. Rather than stating the formula here, we will consider some of the factors that enter into it. First of all, the size of a sample that's required to estimate a population parameter depends on the magnitude of the estimation errors that can be tolerated. The entire population must be sampled if the parameter must be known exactly. If a sample is taken, there will almost always be some estimation error, and for some samples the error may be very large. Since simple random sampling is consistent, the variance of estimation errors can be reduced by increasing the sample size.

Three factors enter the sample size formula for simple random sampling: the size of the population, the variance of the variable that is to be estimated, and the size of the estimation error that can be tolerated. Some rules of thumb for these factors are as follows: The larger the population size, the smaller the percentage that must be sampled in order to realize an estimator variance of a given size. For example, with a population of 100 pupils it might be necessary to sample 50 percent (or 50 out of 100), but with a population of 10,000 pupils it might only be necessary to sample one percent (or 100 out of 10,000) to realize a given estimator variance.

The larger the variance of the variable for which a parameter is to be estimated, the larger the sample size required to achieve a given estimator variance. This is intuitively reasonable. If the variable (for Objective 1, reading achievement) has a large variance, estimates will fluctuate greatly from sample to sample; a larger sample size will be required to reduce its average fluctuations. Finally, the smaller the estimation error that can be tolerated, the larger will be the required sample size. Again, this rule is intuitively reasonable.

Should simple random sampling really be used to achieve Objective 1? Probably not, for the following reasons. First, there are other, more-efficient sampling methods that can be used. Second, it would be administratively cumbersome to use simple random sampling. As previously mentioned, the assessment director would need a complete list of all fifth-graders enrolled in the state. While such a list could probably be compiled in most states, its preexistence is doubtful, and its compilation would be expensive. When sampled fifth-graders were actually tested, some classes of 25 would have 20 tested pupils, some would have only one or two tested pupils, and some would have none at all. Testing only some of the pupils in a classroom is administratively cumbersome, and probably should be avoided unless the number of pupils drawn from each classroom is very small.

Simple random sampling is almost always discussed in sampling texts because it is a straightforward procedure, and can be used to illustrate important sampling properties. It also provides a benchmark against

which the efficiency of more sophisticated sampling procedures can be compared. For statewide assessment the practicality of simple random sampling is limited, although it may be useful when the objective is to estimate some property of schools or school districts.

Stratified Random Sampling

An alternative to simple random sampling that could be used to achieve Objective 1 is stratified random sampling. Stratified random sampling is generally more efficient than simple random sampling, because it takes advantage of facts that are known about the elements of a population. Stratified random sampling can be contrasted with simple random sampling by considering a specific example. Suppose that the size of a simple random sample necessary to estimate the average reading achievement of a state's fifth-graders was found to be 200. Following the procedure for selecting a simple random sample, it is possible that the 200 pupils selected might have an achievement average that was far higher than the average for all fifth-graders in the state. This would almost surely be the case if most of the pupils in the sample had verbal IQ scores that were, say, above 130. Suppose it was possible to guard against samples that had almost all high-IQ pupils, by ensuring that any sample selected would have some low-IQ pupils, some mid-IQ pupils and some high-IQ pupils, with percentages of each similar to the percentages for the whole state. Samples of pupils that came close to representing the state's fifth-graders on verbal IQ would probably do a good job of representing them on reading achievement. This is true because verbal IQ-score and read-

ing achievement are highly related; those with high verbal IQ-scores are likely to have high reading achievement scores, and those with low verbal IQ-scores are likely to have low reading achievement scores. Use of known relationships among variables and available data on sampling units is what makes stratified sampling efficient. Stratified sampling prevents the selection of extremely unrepresentative samples (such as all high-IQ pupils), and thereby prevents large estimation errors. To achieve an estimator variance of a given size, stratified sampling will therefore require a smaller sample size than will simple random sampling.

In stratified random sampling, elements of the population are first classified into categories called strata, according to their values on one or more stratification variables. In the previous example, verbal IQ played the role of a stratification variable. Any variable for which a value is known for every element of the population can be used as a stratification variable. However, stratified sampling, won't be efficient unless the stratification variable and the variable for which estimates are desired (reading achievement in the previous example) are highly related.

Considering the previous example more explicitly, suppose that verbal IQ was to be used as a stratification variable, and the parameter to be estimated was the average reading achievement of all fifth-graders in a state. The first step in using stratified random sampling would be to define appropriate strata. For example, low-IQ pupils might be defined as those with verbal IQ-scores below 85, mid-IQ pupils might be defined

as those with verbal IQ-scores between 86 and 115, and high-IQ pupils as those with verbal IQ-scores of 116 or more. These IQ intervals would define three strata, and might be labeled stratum 1, stratum 2 and stratum 3. Once the strata were defined, each fifth-grader in the state would be classified as a member of stratum 1, 2 or 3 depending on his (her) verbal IQ-score. When all fifth-graders in the state had been assigned to strata, a simple random sample of pupils would be drawn from each stratum. The average reading achievement of pupils sampled from each stratum would then be calculated, and these averages would be weighted appropriately to form an estimate of the average achievement of fifth-graders throughout the state. The estimator would be both unbiased and consistent.

For estimating a statewide average, stratified random sampling has the same disadvantages as simple random sampling. It requires a sampling frame that lists all fifth-graders in the state. In addition, it might result in selection of a few pupils from some classes and many pupils from others. It thus has the potential of being administratively disruptive in some schools and districts.

The main advantage of stratified random sampling is its efficiency (when the right stratification variables are used). In addition, when stratified sampling is used in statewide assessment or in other educational data-collection programs, the information needed for stratification is generally available. During the last decade at least, group IQ testing has been almost universal, and nearly all school districts administer standardized achievement tests (Goslin, 1967). In addition, school systems record all manner of information on their pupils such as parental

occupations, educational levels of parents, and sizes of pupils' families. All of these variables tend to be highly related to current educational achievement (Mollenkopf and Melville, 1956; Burkhead, 1967), and if available, would be quite useful as stratification variables in state-wide assessments.

In theory, strata can be defined by any number of variables. One could, for example, stratify pupils by IQ-score and status-level of father's occupation. The strata thus formed might be labeled low-IQ and low-status occupation, low-IQ and mid-status occupation, low-IQ and high-status occupation, mid-IQ and low-status occupation, etc. Stratification by two or more variables is only efficient when each stratification variable is highly related to the variable for which estimates are sought, and when the stratification variables are not highly related among themselves. The previous example, stratification of pupils by IQ-level and by status level of father's occupation, would probably be an unnecessarily cumbersome procedure. Although reading achievement is highly related to both IQ-level and status-level of father's occupation, the two stratification variables are themselves highly related. Pupils from high-status homes tend to have higher IQ levels, and vice versa. Stratifying pupils by these two variables is therefore redundant; stratification by either variable would be almost as efficient as stratification by both; although IQ-level would probably be a better stratification variable than would father's occupation.

Practical use of stratified sampling requires several design decisions in addition to those already discussed. Once stratification variables have been chosen, the sample designer must decide how many strata to use,

the limits or boundaries for each stratum (e. g., IQ below 90, IQ between 91-110, etc.), the size of the sample to select, and the number of units to sample from each stratum. Each of these topics has been the subject of theoretical and empirical study in the theory of sampling. Again, some practical factors that influence the decisions will be described. The choice of number of strata depends on the magnitude of the relationship between the stratification variable and the variable for which estimates are sought. The stronger the relationship, the larger the number of strata that will prove useful, although practical limits are reached very quickly. Even when the stratification variable and the variable of interest have a correlation coefficient of 0.90, there is not much advantage to using more than four strata (Cochran, 1963). The problem of determining boundaries for strata so as to make stratified sampling as efficient as possible has been given considerable attention by Dalenius and Hodges (1959). They provide formulas that can be used in practice, but defy simple, intuitive explanation. Explicit formulas also exist for determining the sample size to use in stratified sampling. As in simple random sampling, required sample size depends on the population size and the size of the estimation errors one can tolerate. Unlike simple random sampling, the sample size for stratified sampling also depends on how well the population has been stratified. The object of stratification is to form categories, within which sampling units are as nearly alike as possible on the variable of interest. The more nearly this has been accomplished, the smaller will be the sample size required to achieve a given estimator variance. Determination of the number of units to be sampled from each stratum is generally

handled in one of two ways. Using a procedure termed optimal allocation, a specific formula indicates the sample size for each stratum. The advantage of this procedure is that it makes a given stratified sampling procedure as efficient as possible (hence the term optimal). An alternative procedure is termed proportional allocation. With proportional allocation, the size of the sample selected from each stratum is proportional to the number of population elements in the stratum. The advantages of proportional allocation include simplified estimation formulas, and assurance that the stratified sampling procedure will be at least as efficient as simple random sampling.

Systematic Sampling

The average reading achievement of fifth-graders in a state could also be estimated by using a systematic sampling procedure. Several systematic sampling procedures have been developed in the last two decades, but only the one used most widely--linear systematic sampling--will be considered.

Like simple random sampling, linear systematic sampling would require a sampling frame of fifth-grade pupils. Instead of consulting a table of random numbers to determine each sampled pupil, a random number table is consulted only once with linear systematic sampling. The sampling frame of pupils is considered to be an ordered list. The first sampled pupil is selected randomly, and successive pupils are selected at multiples of a constant interval beyond the first. A specific example may help to clarify the procedure.

Suppose it was desired to select a linear systematic sample consisting of ten percent of the fifth-graders in the population. To determine the first sampled pupil, a number between one and ten would be drawn from a random number table. The pupil with the corresponding number on the sampling frame would become the first sampled pupil. Thereafter, every tenth pupil would be sampled. Thus if the random number six were drawn from the table, the first sampled pupil would be the one listed sixth in the frame, the next sampled pupil would be listed 16th in the frame, the next 26th, and so on, until the sampling frame had been exhausted.

NUMERICAL EXAMPLE. Consider the selection of a ten percent systematic sample from a population of fifth-grade pupils. Suppose that a table of random numbers had been consulted to select a number between one and ten, and that the number drawn was six. If the sampling frame were as follows, the sampled pupils would be those marked with an asterisk:

Pupil Number	*Pupil Name*
1	*Murphy, John*
2	*Centra, Paul*
3	*Bruno, Barbara*
4	*Aron, Carol*
5	*Parker, Mary*
**6*	*Nesbitt, William*
7	*Lee, Marjorie*
8	*Sinclair, Susan*
9	*Thomas, George*
10	*Wichert, Jane*
11	*Urban, Paula*
12	*Mann, Marica*

Pupil Number	*Pupil Name*
13	*Tocco, Brenda*
14	*Malcolm, Thomas*
15	*Angoff, Douglas*
**16*	*Fouratt, Sharron*
17	*Brambley, Joan*
18	*Willis, Kevin*
19	*Picard, Ronald*
20	*Libby, Linda*
21	*Arcieri, Sheryl*
22	*Kristof, Charles*
23	*Patterson, Virginia*
24	*Johnson, Elmer*
25	*Saxe, Anne*
**26*	*Stahl, Mildred*
27	*Walsh, Helen*
28	*Adams, Patricia*
.	
.	
.	

The three dots signify the continuation of the list, and the selection of every tenth pupil beyond the 26th, until the entire sampling frame had been exhausted. Thus if the list contained 240 pupils, the last one selected for the sample would be pupil number 236.

Systematic sampling has the advantage that it is easy to apply by hand, whereas simple random sampling or stratified random sampling are quite tedious without a computer when a sample of appreciable size must be drawn. When used in an assessment program, systematic sampling would

also ensure that the number of pupils sampled from each classroom was approximately equal, provided the sampling frame listed pupils sequentially by classroom. Like simple random sampling though, systematic sampling would require a list of all fifth-graders in the state.

Unlike simple random sampling and stratified sampling, linear systematic sampling is sometimes undependable. It is not always consistent, and there are no really good ways to estimate mean square error. Conversely, linear systematic sampling can be very efficient if the list used for sampling is carefully constructed. If pupils were listed alphabetically in the sampling frame, one would suppose that their average achievement might be estimated about as efficiently as with simple random sampling. In fact, alphabetic listing of pupils sometimes results in more efficient estimation (Jaeger, 1970), although this won't always be the case. Real gains in the efficiency of systematic sampling can be realized by listing pupils in increasing order on some variable that is highly related to the variable of interest. For example, if a linear systematic sample of fifth-graders was selected from a sampling frame in which pupils were listed in increasing order of their verbal IQ-scores, average reading achievement could be estimated very effficiently. The effect of such ordered listings is much the same as the effect of stratification, since sampling from an ordered list ensures that some pupils are sampled at all levels of the variable used for ordering.

Linear systematic sampling is one of those procedures, mentioned earlier, that isn't always consistent and, depending on the relationship between the sample size and the population size, may lead to biased esti-

mation. Usually the magnitude of the estimation bias is inconsequential, but the lack of consistency may prove to be a serious problem. If sampling must be done without a computer and if the required sample size is large, linear systematic sampling should be considered for statewide assessment. Otherwise, alternative sampling procedures (such as stratified sampling) will provide more dependable results.

Cluster Sampling

In the sampling procedures discussed to this point, the sampling units used were basic elements of a population; e. g., individual pupils. In cluster sampling, the sampling units are not basic population elements but are groups or aggregations of such elements. These groups of elements are termed clusters.

In most applications of cluster sampling, the clusters used are naturally-occurring groups. In surveys of consumer behavior, for example, homes are frequently used as sampling units. When estimating the average achievement of fifth-graders throughout a state, several naturally-occurring clusters of pupils might be used--school districts, schools, or homerooms. Of course, these aren't the only possibilities for clusters. One might consider groups of students living in particular areas of the state or groups of pupils with last names beginning with the same letter. However, naturally-occurring clusters afford far greater administrative convenience than would these contrived clusters. Pupils can readily be identified by classroom, school or school district, and could easily be assembled for testing and measurement on a homeroom-by-homeroom or school-by-school basis.

If a cluster sampling procedure is identified by the units used as clusters--school districts, schools, homerooms, or combinations of these--many different cluster sampling procedures could be used to gather data for Objective 1. Before enumerating some of the possibilities, let's consider one in detail, and thereby introduce some of the language of cluster sampling.

Suppose it was decided to use schools as clusters, and to test the reading achievement of all fifth-graders enrolled in sampled schools. This procedure is an example of single-stage cluster sampling. The sampling plan would be carried out by first constructing a sampling frame of all schools in the state that enrolled fifth-grade pupils. A simple random sample of schools could then be selected using a table of random numbers, just as in simple random sampling of pupils, described above. All of the fifth-grade pupils in sampled schools would then be given a reading achievement test, and appropriate formulas would be applied to the test results in order to estimate average achievement for the state. The formulas to be used (estimators) are well known in the sampling theory literature, and can be found in any standard text such as Cochran (1963).

This cluster sampling procedure has some obvious administrative advantages. First, the state department of education is likely to have a complete list of schools that enroll fifth-graders, although it probably doesn't have a list of fifth-graders enrolled in the state. Thus a ready-made sampling frame is likely to exist for this sampling procedure. Second, only a sample of schools will be involved in testing. Disruption

of normal academic procedures will be confined to the sample of schools, the costs of distributing testing materials will be reduced, and administrative procedures will be simplified.

The administrative convenience of this sampling procedure is likely to be offset by a substantial reduction in efficiency. In almost all cases, cluster sampling of schools will be far less efficient than simple random sampling of pupils. The "almost" is inserted in the previous sentence because there are notable exceptions to the rule. The efficiency of single-stage cluster sampling depends on many factors, some of which can be controlled by the sample designer. The composition of the clusters used influences efficiency to a large degree. Two extreme cases will illustrate this point. To take one extreme, suppose that all of the fifth-graders in any given school had the same reading achievement score. In this case, testing all the fifth-graders in a school would be a waste of time and money; the average achievement in a school could be determined by testing just one fifth-grader. More to the point, the effective sample size is equal to the number of schools in which testing takes place, rather than the number of pupils tested (since testing more than one pupil in a school would provide only redundant information). In technical terms, this extreme case represents a situation in which all of the elements within a cluster are completely homogeneous on the variable to be estimated. The other extreme would occur in a situation where the average reading achievement of fifth-graders in each school was identical, and equalled the average for the whole state. In this

case, the average for the state could be estimated perfectly by collecting data in only one school, since testing pupils in more than one school would provide only redundant information. In technical terms, this extreme represents a situation in which elements within a cluster are as heterogeneous as elements within the entire population, and where clusters are completely homogeneous. In real life, the composition of the population will fall somewhere between these extremes. For cluster sampling to be efficient, we would like the composition of the population to be similar to the second extreme: not much difference among clusters on the variable to be estimated, and a lot of heterogeneity among elements in the same cluster. With this composition, only a few clusters need be sampled in order to get a good representation of the entire population.

Unfortunately, the naturally-occurring clusters available for statewide assessments tend to provide homogeneity within clusters and heterogeneity between clusters for many variables likely to be of interest. Consider sampling of schools to estimate pupil achievement. At least before bussing for purposes of desegregation, the attendance areas of schools tended to be defined by neighborhoods that were relatively homogeneous in their socio-economic and racial compositions. In a society where neighborhoods tend to be defined by people of the same social and economic level, it is natural that schools tend to be homogeneous in these variables. Since pupils' scores on achievement tests are highly related to the socio-economic

status of their families, schools also tend to be homogeneous in measured academic achievement.

The composition of the population of interest (e. g., all fifth-graders in a state) is a factor beyond the control of the sample designer; whatever is found must be tolerated. However, there are factors that the user of cluster sampling can control so as to greatly increase sampling efficiency. One such factor is the estimation procedure employed. When the clusters to be sampled are not only heterogeneous, but also tend to vary greatly in size (both are tendencies of schools and school districts), simple random sampling of clusters with unbiased estimation of averages is very inefficient. A more efficient alternative involves simple random sampling of clusters and use of an estimation procedure known as ratio estimation. To use ratio estimation, the number of elements in each cluster must be known; a requirement that is easily met in most assessment applications. The ratio estimator is biased, but consistent. The amount of bias is likely to be small for populations used in statewide assessments, and the mean square error will usually be much smaller than that of the unbiased estimator. Formulas for ratio estimation can be found in Murthy (1967) Cochran (1963) and Hansen, Hurwitz and Madow (1953).

Additional alternatives modify both the sampling procedure and the estimation procedure used with single-stage cluster sampling. By definition, each cluster has an equal chance of being selected when clusters are sampled randomly. One alternative procedure, known as PPS sampling, selects clusters with probabilities proportional to their sizes. If schools were

being used as clusters in order to estimate average fifth-grade reading achievement, the probability of selecting a given school would depend on its fifth-grade enrollment. A school with 200 fifth-graders would be twice as likely to enter the sample as would a school with 100 fifth-graders. The PPS procedure provides not only a sampling method but associated estimators of averages, proportions and variances as well. It is simplest to do PPS sampling "with replacement" since selection probabilities vary as the sample is drawn, when sampling is done without replacement. PPS sampling with replacement provides unbiased estimation, but is an inconsistent procedure. The mean square error of the estimator gets consistently smaller as sample size is increased, but does not go to zero when the sample size equals the population size. In practical situations, this lack of consistency will be a problem only when the required sample size is very close to the population size.

PPS sampling is efficient only when cluster size is highly related to the variable for which estimates are desired. Since school size and school district size are not highly related to basic-skills achievement (Burkhead, 1967), PPS sampling will not be efficient for estimation of average achievement in a state. Some school and district "input" variables (such as the average value of the taxable property in an attendance area or district) are highly related to school or district size, and PPS sampling would probably be very efficient for estimation of these variables.

A final alternative, PPES sampling, is likely to be a very efficient way of estimating average achievement in a state. PPES stands for "prob-

ability proportional to expected size" (Cochran, 1963), a term that is appropriate in some sampling contexts but not in the context of statewide assessment. PPES sampling was first introduced to handle situations in which cluster sizes were not known exactly. In these cases, "expected sizes" rather than actual sizes were used.

In assessment applications, cluster sizes are usually known but are often nearly unrelated to the variables for which estimates are desired. The greater the relationship between the variable for which estimates are sought and the "expected size" variable, the higher the efficiency of PPES sampling. This being true, clusters can be sampled with probabilities proportional to any variable that has a known value for every cluster in the population; the variable used can be totally unrelated to cluster size. Consider the case of Objective 1. Suppose that a group IQ-test had been administered to every fourth-grader in the state in the year preceding the current assessment. If the state had records containing the average IQ of fourth-graders for each school and the fourth-grade enrollment of each school, the product of these two could be used very effectively as an "expected size" measure when estimating average fifth-grade reading achievement. This procedure would be highly efficient because the average of fourth-grade IQ-scores and the average of fifth-grade reading achievement scores would be highly related across schools.

Like PPS sampling, PPES sampling results in unbiased but inconsistent estimation. Again, inconsistency will be a practical problem only when the required sample size is very close to the population size. Additional information on PPS sampling and PPES sampling can be found in Murthy (1967) and in Cochran (1963).

Instead of using schools as clusters, the average reading achievement of fifth-graders in the state could be estimated by using either homerooms or school districts as clusters. Either of these single-stage cluster sampling procedures would be feasible, provided appropriate sampling frames could be constructed. Undoubtedly, every state department of education has a complete listing of school districts that enroll fifth-graders. A sampling frame of homerooms probably wouldn't exist in most states though, and sampling by homerooms would require a specially constructed frame. The cost of constructing a sampling frame of homerooms would probably be more than offset by the increased efficiency of a single-stage cluster sampling plan with homerooms as clusters. In most states, cluster sampling of homerooms would be far more efficient than cluster sampling of schools, and cluster sampling of schools would be more efficient than cluster sampling of districts. The increased efficiency is due in part to substantially greater size variability among districts than among schools, and among schools than among homerooms.

Thus far we have considered only single-stage cluster sampling procedures. Many multi-stage cluster sampling procedures could be used to estimate the average reading achievement of a state's fifth-graders. Possibilities include the following: 1) A random sample of schools could be drawn, and within sampled schools, random samples of homerooms could be selected. All fifth-graders in sampled homerooms would be tested. 2) A random sample of districts could be drawn, and within sampled districts,

random samples of schools could be selected. All fifth-graders in sampled schools would be tested. 3) A random sample of districts could be drawn, and within sampled districts, a random sample of homerooms could be selected. All fifth-graders within sampled homerooms would be tested. 4) A random sample of districts could be selected, and within sampled districts, random samples of fifth-graders could be selected and tested. 5) A random sample of schools could be drawn and within sampled schools, random samples of fifth-graders could be selected and tested. 6) A random sample of fifth-grade homerooms could be selected, and within sampled homerooms, random samples of pupils could be drawn and tested. 7) A random sample of districts could be selected, random samples of schools could be drawn within sampled districts, and random samples of homerooms could be selected within each sampled school. All fifth-grade pupils within sampled homerooms would be tested. 8) A random sample of districts could be selected, random samples of schools could be drawn within sampled districts, random samples of homerooms could be drawn within sampled schools, and random samples of pupils would be selected and tested within sampled homerooms. Although these eight procedures do not exhaust the possibilities, they provide sufficient illustration of the flexibility of cluster sampling.

Procedures 1) through 6) are examples of two-stage cluster sampling. In procedure 2), for example, sampling of districts constitutes the first stage (districts are termed primary sampling units or PSU's), and sampling of schools is the second stage. Schools would be called secondary sampling units. Procedure 7) is an example of a three-stage cluster sampling procedure, with districts as PSU's, schools as secondary sampling units, and

homerooms as tertiary sampling units. Procedure 8) is a four-stage cluster sampling procedure.

Multi-stage cluster sampling will often be more statistically efficient than single-stage cluster sampling. That is, the mean square error of the estimator will be smaller, for a given number of elementary units in the sample. There are also some administrative advantages to multi-stage sampling. If sampling frames don't exist, they need only be constructed for a sample of PSU's. For example, if a state wanted to use homerooms as clusters but didn't have the required sampling frame, it could use two-stage sampling with districts as PSU's and homerooms as secondary sampling units. The district sample would be chosen first, and sampling frames of homerooms would be needed only for sampled districts.

Cluster sampling can also be used in combination with other procedures such as stratified sampling or systematic sampling. One could, for example, select samples of schools stratified by the average IQ-level of enrolled fifth-graders or by a measure of the average socio-economic status of pupils' families. As another alternative, one could select a simple random sample of school districts, and select systematic samples of fifth-graders from lists arranged in order of increasing IQ-score within each sampled district. Each of these alternatives would be more efficient than multi-stage random sampling.

The final choice among cluster sampling procedures depends on many factors, not the least of which is previous knowledge of the population of interest. To choose among sampling procedures intelligently, one should have some idea of the degree of homogeneity within and among

potential clusters, and the relationships among variables for which estimates are sought and those that might be used for stratification or as measures of size. Even with these kinds of data, assurance that one has chosen the best of the available alternatives can only come through careful analysis and often, lengthy computation. (See Appendix A).

It cannot be overemphasized that data typically available in schools and school districts can be used very effectively to design efficient sampling procedures. A wealth of information on students, teachers, classes, schools and school districts is routinely recorded and filed in school district offices and in offices of state departments of education. Data from previous testing programs are abundantly available in almost all school districts and states. Background information on pupils and teachers is also on file in most school districts. If judiciously selected and evaluated, these data can be used for stratification, for arrangement of populations in ordered lists, and for pretesting of potentially efficient sampling procedures. This mechanical use of information to arrange and sort populations should not provoke charges of invasion of privacy, since individuals' names need be associated with individual data elements only for purposes of sampling.

Matrix Sampling

Each of the sampling procedures considered to this point has assumed that all sampled pupils respond to the same set of measures; e. g., the same reading comprehension test. In the past ten years, researchers have paid increasing attention to procedures that sample test items as well as students. These procedures are termed multiple matrix sampling,

and have been used successfully in National Assessment as well as in several statewide assessments.

Multiple matrix sampling could be used to estimate the average reading achievement of all fifth-graders in a state. The procedure might be as follows. Suppose that a 50-item reading achievement test was to be used. Instead of administering the entire test to all sampled pupils, the test could be divided into five forms with ten items each. Each sampled pupil would then take a 10-item form instead of the entire 50-item test. Each of the 50 items would be used in a 10-item form, and approximately equal numbers of pupils would complete each 10-item form. Lord (1955; 1962) has developed formulas for estimating the average score pupils would have earned, had each completed the entire 50-item test. Empirical studies of the best way to divide tests into forms and the sizes of pupil samples to use with each form have been completed by Shoemaker (1970; 1971) and Knapp (1968), among others.

To date, statistical procedures for analysis of multiple matrix sampling have been developed only for simple random sampling of items and pupils. Although more complex designs can be used, needed analytic procedures are not yet available.

Objective 2: Estimating the proportion of third-graders in each school district who can successfully achieve an arithmetic objective.

Some statewide assessments use test items that are specifically designed to measure the achievement of particular objectives. For example, an assessment might include items designed to measure achievement of the arithmetic objective "Addition of pairs of single-digit integers".

Five such items might be administered to a pupil, and the pupil might be said to have achieved the objective provided he can successfully complete three of the five items.

Suppose that a statewide assessment contained such objectives-related items, and that the principal purpose of the assessment was to determine the proportion of pupils in each of the state's school districts that had achieved each designated objective.

Many of the sampling procedures described above could be used to achieve Objective 2. Only in very small school districts (e. g., those with grade three enrollments under 200) would sampling be uneconomical. Among the procedures that might be used to achieve Objective 2 are simple random sampling of pupils, stratified random sampling, linear systematic sampling, and some forms of cluster sampling.

With Objective 2, each school district's third-graders would constitute a separate population, and sampling in each school district could be handled differently. That is, one district might use simple random sampling, while another might use two-stage cluster sampling of schools and homerooms, with homerooms stratified by average ability level of pupils. In practice, use of several different sampling procedures would make good sense if the districts varied greatly in size. While cluster sampling would be infeasible in a small school district (say, one with only three elementary schools), it might prove to be highly efficient in a state's largest school districts.

To accomplish Objective 2, simple random sampling would be handled just as it is described for Objective 1. Standard formulas exist for

the estimation of proportions through simple random sampling, as they do for the estimation of mean square errors (Murthy, 1967; Hansen, Hurwitz and Madow, 1953).

When the objective is estimation of a proportion, stratified sampling is unlikely to afford appreciable increases in efficiency over simple random sampling. To be efficient, stratified sampling requires that variances within strata be much smaller than the variance within the whole population. The variances of proportions are very similar, unless the proportions are extremely large or extremely small (the variances of proportions in the range 0.2 to 0.8 are very similar). Thus little reduction in the variance of proportions can be gained from stratification.

Use of linear systematic sampling is just as reasonable for the achievement of Objective 2 as it was for the achievement of Objective 1. The same potential advantages, and the same cautions, apply. A school district is more likely than a state department of education to have past test data and other information on individual students. This information can be used to create ordered sampling frames, permitting systematic sampling from an ordered list.

Unless a school district is very large, multi-stage cluster sampling will not be practical. For moderately large school systems (enrollments of ten thousand to thirty thousand), single-stage cluster sampling of homerooms is likely to be administratively practical and statistically efficient for estimation of averages or proportions. Compiling a list of third-grade homerooms should not be difficult in a district of moderate

size. Sampling by homeroom would permit testing of intact groups of pupils, and would provide a convenient route for distribution of materials and handling of assessment materials in the field.

Multiple matrix sampling could also be economical and convenient in all but the smallest school systems. Shoemaker (1970) has shown that multiple matrix sampling is useful for estimation of averages, provided the population is no smaller than 300.

Summary

This paper was intended to help the reader become conversant with important sampling terms and concepts, and to become aware of sampling procedures that might be used in a statewide assessment. It was not intended to create instant sample-design experts or sampling theorists.

If the reader has gained a basic understanding of such terms and concepts as estimate, estimator, population parameter, estimator bias, etc., and if some of the sampling options available for statewide assessments are now intelligible, the paper has accomplished its purpose.

Designing an efficient sample requires knowledge of the science of sampling. But perhaps more than in other statistically-oriented disciplines, good sample design is an art. It requires a sensitivity to the nature of the populations of interest, and attention to information and data that might, to the novice, seem unrelated to the sampling task at hand. For these reasons, there is no substitute for experience when a truly efficient sample design is desired. Investment in expert sampling consultation will usually be repaid many times over

by the economies an efficient design provides. But it behooves the assessment director to be conversant, if not expert, on sampling and its potentials. By knowing a little about the subject, the right questions can be asked, and the right data can be provided. The task of the sample designer will be made easier, and the resulting product all the better.

REFERENCES

Burkhead, J., Input and Output in Large City High Schools. New York: Syracuse University Press, 1967.

Cochran, W. G., Sampling Techniques. New York: John Wiley and Sons, 1963.

Dalenius, T. and J. L. Hodges, Jr., "Minimum variance stratification", Journal of the American Statistical Association, 54 (1959), 88-101.

Goslin, D., Teachers and Testing. New York: Russell Sage Foundation, 1967.

Hansen, M., W. Hurwitz and W. G. Madow, Sample Survey Methods and Theory. 2 vols. New York: John Wiley and Sons, 1953.

Jaeger, R. M., "Designing school testing programs for institutional appraisal: an application of sampling theory", Stanford, California: 1970. Unpublished doctoral dissertation.

Knapp, Thomas R., "An application of balanced incomplete block designs to the estimation of test norms", Educational and Psychological Measurement, 28, (1968), 265-272.

Lord, F. M., "Sampling fluctuations resulting from the sampling of test items", Psychometrika, 20 (1955), 1-23.

Lord, F. M., "Estimating norms by item sampling", Educational and Psychological Measurement, 22, (1962), 259-267.

Mollenkopf, W. G. and S. D. Melville, "A study of secondary school characteristics as related to test scores", Educational Testing Service, Research Bulletin 56-6. Princeton: 1956.

Murthy, M. N., Sampling Theory and Methods. Calcutta: Statistical Publishing Society, 1967.

Shoemaker, D. M., "Allocation of items and examinees in estimating a norm distribution by item-sampling", Journal of Educational Measurement, 7, (1970), 123-128.

Shoemaker, D. M., "Further results on the standard errors of estimate associated with item-examinee sampling procedures," Journal of Educational Measurement, 8, (1971), 215-220.

APPENDIX A

Evaluation of Alternative Cluster Sampling Procedures--An Example

When choosing among alternative cluster sampling procedures, the kinds of theoretical notions discussed in this paper (a procedure will be more efficient when cluster sizes don't vary much, heterogeneity within clusters and homogeneity between clusters will provide increased efficiency, etc.) provide some guidance. In a specific application, assurance that one is using the best procedure can also be gained through analysis of data from the school district or state where sampling is to be used.

Many characteristics of schools, school districts, and groups of students show remarkable stability from year to year. For example, the average basic skills achievement of a school's fourth-grade class is likely to be very similar in two successive years, as is the socio-economic composition of the school's student body. When searching for a sampling procedure that provides maximum efficiency, one can take advantage of this kind of stability. The method is as follows: Use data from the previous school year to evaluate the efficiency of the sampling procedures being considered for the current year. Since it is unlikely that sampling has been used in the past, data will be available for all students, classes and schools in the district or state. With data available for the entire population (a situation that will <u>not</u> hold for the current school year if sampling is used), results of sampling the previous year's population using a variety of procedures can be readily compared.

An example of this kind of evaluation uses data from a single school district, called Anydistrict (Jaeger, 1970). For simplicity, computation of estimates and estimator variances will not be shown; only initial data and final results will be presented.

The population parameter to be estimated in this example is the average reading achievement of the district's sixth-graders. The sixth-grade enrollment of the district is 1180, with 45 sixth-grade classes in 21 schools. Data available from the previous school year include the average sixth-grade reading achievement in each school, the sixth-grade enrollment in each school, and the average verbal ability score of fifth-graders in each school. These data will be used to evaluate four alternative cluster sampling and estimation procedures: Simple random sampling of schools with unbiased estimation, simple random sampling of schools with ratio estimation, sampling of schools with probabilities proportional to their sixth-grade enrollments (PPS sampling and estimation), and sampling of schools with probabilities proportional to totals of fifth-grade ability test scores (PPFS sampling and estimation).

The evaluation of each cluster sampling procedure will use data from the entire population of 21 schools. With these data, estimator variances can be calculated exactly. It must be emphasized that data for the entire population will be available only when all sixth-graders in the district are tested--a situation that will not obtain in the current school year, when sampling is used. The method then, is to use population data from a previous school year to evaluate alternative sampling procedures, and to assume that the most efficient procedure for one school year will also be most efficient for the next year. The assumption is generally sound.

The following table shows sixth-grade average reading achievement scores, sixth-grade enrollments, and average fifth-grade ability test scores for the 21 schools in the district under study. The data are real. They were provided by the research office of a medium-sized school district.

Table A: Sixth-grade Average Reading Achievements, Sixth-Grade Enrollments, and Average Fifth-Grade Ability-Test Scores for Elementary Schools in Anydistrict.

School Number	Average Grade 6 Reading Achievement*	Grade 6 Enrollment	Average Grade 5 Ability Score
1	66.11	56	33.54
2	66.83	65	32.96
3	71.27	71	38.06
4	56.09	58	33.81
5	64.57	47	34.29
6	71.09	66	37.84
7	74.89	55	36.70
8	70.67	99	37.69
9	74.51	57	39.06
10	68.13	40	37.19
11	70.02	59	36.10
12	72.57	72	39.90
13	58.86	43	35.36
14	66.35	63	36.20
15	70.71	38	36.92
16	65.82	51	34.42
17	70.98	51	35.15
18	67.56	41	33.51
19	82.21	29	40.76
20	65.61	74	35.02
21	51.14	49	30.18

*Average number of test items correct.

The data in Table A were used in formulas for the variance of the estimated mean, appropriate to each of the four cluster sampling and estimation procedures. In all cases, it was assumed that 10 of the 21 schools in Anydistrict were sampled, and that all sixth-graders in sampled schools were tested. The sampling and estimation procedure that provided the smallest variance was judged to be best.

To evaluate PPS sampling, it was assumed that schools were sampled with probabilities proportional to their sixth-grade enrollments (the data in the third column of Table A). To evaluate PPES sampling, a slightly more-complex assumption was made. The measure of "size" used for a school was equal to the product of the school's sixth-grade enrollment, and the average ability-test score earned by the school's fifth-graders (the data in columns three and four in Table A). While this product (sixth-grade enrollment times fifth-grade ability test score) might not have much meaning as an assessment statistic, it makes an excellent variable for PPES sampling since it is highly correlated with the total of sixth-grade reading achievement scores in a school.

The variances of estimators of average sixth-grade achievement in the district are given in Table B, below:

Table B: Variances of Estimators of Average Achievement for Sixth-Grade Students in Anydistrict. Sample Size is 10 Schools from a Population of 21.

Sampling and Estimation Method	Estimator Variance
Simple random sampling of schools with unbiased estimation	21.790
Simple random sampling of schools with ratio estimation	1.802
Sampling of schools with probabilities proportional to sixth-grade enrollments (PPS)	3.622
Sampling of schools with probabilities proportional to fifth-grade ability test scores (PPES)	1.358

From the data in Table B, it is clear that PPES sampling of schools is the most efficient of the four cluster sampling procedures. PPES sampling is slightly more efficient than simple random sampling of schools with ratio estimation, more than twice as efficient as PPS sampling of schools, and more than sixteen times as efficient as simple random sampling of schools with unbiased estimation. Efficiency is calculated from the ratio of estimator variances.

Although PPS sampling and PPES sampling are not consistent procedures, the variances of their estimators do decrease steadily as sample size is increased. Simple random sampling of clusters with unbiased estimation or with ratio estimation are consistent, so the variances of their estimators also become steadily smaller as sample size is increased. Thus one can generalize from the data in Table B for all sample sizes that are substantially smaller than the population size. PPES sampling will be most efficient, simple random sampling of schools with ratio estimation will be next most efficient, PPS sampling will rank third in efficiency, and simple random sampling of schools with unbiased estimation will be very inefficient.

The formulas used to calculate estimator variances in this example can be found in many sampling texts, including Murthy (1967), Cochran (1963) and Hansen, Hurwitz and Madow (1953).